Fusing highly dimensional energy and connectivity features to identify affective states from EEG signals

Arnau-González, Pablo; Arevalillo-Herráez, Miguel; Ramzan, Naeem

Published in:
Neurocomputing

DOI:
10.1016/j.neucom.2017.03.027

E-pub ahead of print: 18/03/2017

Document Version
Peer reviewed version

Fusing highly dimensional energy and connectivity features to identify affective states from EEG signals

Pablo Arnau-González[a,*], Miguel Arevalillo-Herráez[b], Naeem Ramzan[a]

[a] *School of Engineering and Computing, University of the West of Scotland, United Kingdom*
[b] *Departament d'Informàtica, Universitat de València, Spain*

Abstract

In this paper, a novel method for affect detection is presented. The method combines both connectivity-based and channel-based features with a selection method that considerably reduces the dimensionality of the data and allows for an efficient classification. In particular, the Relative Energy (RE) and its logarithm in the spacial domain, and the spectral power (SP) in the frequency domain are computed for the four typical frequency bands (α, β, γ and θ), and complemented with the mutual information measured over all channel pairs. The resulting features are then reduced by using a hybrid method that combines supervised and unsupervised feature selection. First, a Welch's t-test is used to select the features that best separate the classes, and discard the ones that are less useful for classification. To this end, all features where the t-test yields a p-value above a threshold are eliminated. The remaining ones are further reduced by using Principal Component Analysis. Detection results are compared to state-of-the-art methods on DEAP, a database for emotion analysis composed of labeled recordings from 32 subjects while watching 40 music videos. The effect of using different classifiers is also evaluated, and a significant improvement is observed in all cases.

Keywords: EEG, connectivity features, energy features, emotion recognition, feature reduction, feature extraction

1. Introduction

To endow computers with the ability to successfully infer or respond to affect, it is necessary to combine research results in diverse areas, which include computer sciences, signal processing and cognitive sciences [1]. The interpretation of affect on bio-signals could lead computers to be affect-responsive, enhancing the user's experience by adapting feedback and modifying the behavior of applications in real-time.

The first step to be able to respond to emotions is affect recognition, that focuses on identifying emotions and other affective phenomena on the subject. The evaluation of the affective state is usually done according to an emotional model that suits the particular application. One of the simplest models is the one described by Ekman, which is composed of six discrete primitive emotions, namely *anger, fear, sadness, surprise, disgust and happiness* [2]. Other alternative models include Plutchik's Wheel of Emotion [3], and Russell's Circumplex Model [4], which locates emotions in a 2D space defined by the arousal (or activation) and valence (or positiveness). The latter was extended in [5] by adding a third dimension (dominance) to avoid overlapping of certain emotions.

In general, these models are used to build a classification scheme that uses features as an input, and yield a prediction related to the user's emotional state as an output. Features can be of a very diverse nature, but one major factor that affects the system's performance is related to the existing implicit relations between the selected features and the user's reaction to changes in the variables considered in the emotional model. Many research works have measured and investigated subject's reactions by using biological signals [6]. These signals include Electroencephalography (EEG), an electrophysiological monitoring method that uses multiple electrodes placed on the scalp to measure voltage fluctuations that result from ionic current flows within the neurons of the brain.

From a classification perspective, a shared difficulty among related research on affect recognition is the relatively low number of samples available for training. This fact restricts the use of high dimensional models, that in the case of EEG-based systems is usually proportional to the number of channels recorded by EEG. One line of work has focused on the use of feature reduction and feature selection methods [7, 8, 9]. Nevertheless, it is common to find in the literature models with more than 50 dimensions for 40 samples [7, 10]. Another line of work has concentrated on using different features. For example, Chen [11] demonstrated that connectivity features can also be used to detect the affective state, at a reasonable level of accuracy.

In this paper, we build on a preliminary version of this work reported in [12], and present a low-dimensional classification scheme that combines a number of features of different nature, namely channel-based (including both time-domain and frequency-domain) and connectivity features. The method relies on a novel adaptive feature reduction scheme that integrates a supervised feature selection mechanism based on a Welch's t-test with a Principal Component Analysis (PCA), to yield low-

*Corresponding author
Email addresses: pablo.arnaugonzalez@uws.ac.uk (Pablo Arnau-González), miguel.arevalillo@uv.es (Miguel Arevalillo-Herráez), naeem.ramzan@uws.ac.uk (Naeem Ramzan)

dimensional data which is fed into a classifier. The method has been exhaustively tested and compared to existing approaches on the DEAP repository [9], a common benchmark for this type of applications. Results show a significant improvement, both in terms of accuracy and weighted F1-score. The contribution is twofold. On the one hand, channel-based and connectivity features are simultaneously used, to yield a more adequate representation that captures different emotional aspects of the EEG signals and better correlates with the labels of interest. On the other hand, an *adhoc* feature reduction method is proposed to cope with the highly dimensional data than results from the combination. The number of dimensions to retain is considered a model parameter and it is decided dynamically by using cross-validation, or integrated into the grid-search when the classifier requires additional parameters. Together, they achieve better results at predicting valence and activation levels, under a typical two-class classification setting commonly employed in the literature, e.g. [9, 11].

The remainder of this paper is structured as follows. In section 2, we describe some of the most relevant related work. Section 3 presents the proposed method in detail. This includes the EEG features, the dimensionality reduction scheme and the classification mechanism that have been used. Section 4 explains the experimental setting that has been used to validate the proposal, and presents the results of an extensive performance comparison to evaluate the gains achieved. Finally, the paper ends with Section 5, in which some conclusions are drawn and further work is briefly explained.

2. Related work

Most previous work on emotion recognition from EEG signals use a typical classification framework. Under this scheme, EEG signals are recorded during specific emotional situations, on a setting that appropriately represents the detection context. The resulting signals are then pre-processed using spatiotemporal filtering and noise reduction methods, to abate artifacts and enhance the signal-to noise power ratio (SNR). Relevant features are then extracted, and used to train a classifier with data that has been labeled according to an specific emotional model. The resulting model is used in production to estimate the most likely emotional state.

The relationship between EEG signals and affective states has been widely studied in the literature, concluding that feature selection significantly affects the classification performance [6, 13]. In this direction, the recent survey in [6] revisits a number of neuropsychological studies that have reported EEG features that correlate with emotions. Another also recent paper [13] presents a review of the most common features and their weighted correlations. In this work, authors study and classify features according to their domain: time domain, frequency-domain, time-frequency domain and electrode combinations. In addition, they identify Hilbert Huang Spectrum (HHS), Higher Order Crossings (HOC) and Higher Order Spectra (HOS) to be correlated with affective levels and to overperform spectral power bands.

In general, EEG features can be classified into channel-based [7, 9, 14] and connectivity features [11, 15]. Channel-based refer to EEG characteristics that are measured at the single-electrode level, considering the EEG activity of each channel separately. On the contrary, connectivity features are based on the functional connectivity between EEG sensors, including relations such as correlation, coherence, differential asymmetry, rational asymmetry, and phase synchronization [6, 13, 16, 17].

The first works in emotion recognition concentrated on using channel-based features exclusively. For example, the use of Higher Order Crossings (HOC) was explored in [18], both using a single-channel and combining several ones. *Liu et al.* [8] proposed a different approach, based on their observation that higher levels of arousal are usually related to higher values of the Fractal Dimension; as much as valence levels relate to fractal dimension differences between concrete electrodes located in the right and left hemisphere. This initial work was validated with their own data set, and later extended in [7] by using Higher Order Crossings [19] and features from the General Higuchi Fractal Dimension Spectra [20] to understand EEGs as multifractal signals. Recently in [21] first use of recurrent neural networks using reservoir computing techniques have shown promising results in Valence levels estimation.

Other recent works have focused on connectivity features. For example, *Chen et al.* [11] have recently studied the performance of a diversity of such features, namely the Pearson correlation connectivity [22], phase coherence [23, 24] and mutual information, which led to the best results. In addition, *Gupta et al.* [15] also used connectivity features on the DEAP dataset [9]. In this case, they employed graph-related features to represent functional connectivity patterns.

Despite that many existing works in the literature use one or another kind, additional gains can be achieved by appropriately fusing both types of features under the same classification framework. In this direction, the method presented as part of the public dataset DEAP [9] combines the two kind of features. In this work, authors used the Spectral Power and Spectral Power Asymmetry of the single channels and the Spectral Power Asymmetry from 14 pairs of electrodes (SPD), then applied a filter with Fisher Discriminant Analysis with threshold 0.3 in order to avoid irrelevant features for the classification step.

Alternatively multimodal approaches have also been used in research, in DEAP [9] Electroencephalogram and Electrocardiogram signals were used in order to classify the emotion. In [25, 26] visual features were used together with EEG-features. These approaches have shown to improve the results of the single modal approaches, mostly for effect of the other complementary data, like facial video, that, generally performs better than the EEG.

Other authors have also turned the classification framework into a regression setting, by considering valence and activation as continuous variables. In [27], the MAHNOB-HCI dataset is used to extract the spectral power density and asymmetry from 14 pre-selected electrodes from windows of 1 second length, with 50% overlapping. Then they applied diverse re-

gression techniques, multilinear regression, support vector regression, conditional continuous random fields (CCRF) [28] and long short-term memory recurrent neural networks (LSTM-RNN) [29]. They also extracted face features from videos, using distances between face points and their first derivative, and adopted two different strategies for fusing them: Feature Level Fusion and Decision Level Fusion. This work has recently been expanded in [30]. Major changes include using the whole set of 32 electrodes, video, and eye gaze data. They report results that show a weak correlation ($\bar{\rho} = 0.33 \pm 0.38$ for LSTM-RNN) between the features used and the affective levels, but there is also a enormous variance that reinforces the assumption that features are subject-dependent.

3. Proposed Method

One of the major factors that affect classification performance is related to the selection of adequate features that maximize the separation of the different classes. Indeed, more information is fed into the classification process by including both connectivity and channel-based features. However, the simultaneous use of both types of features causes an increase in the dimensionality of the data, that can potentially cancel the positive impact of the information increase. Despite that the Bayes error approaches zero as the number of dimensions increase, it approaches 0.5 when the classifier is trained from a finite number of samples [31]. For this reason, the fusion of connectivity and channel-based features need to be combined with a convenient dimensionality reduction scheme that adapts to the particular problem and retains a number of features that falls within the dimensionality interval of near optimal performance.

With these issues in mind, we have fused connectivity and channel-based features within a typical classification scheme, and used a combination of supervised and unsupervised feature reduction methods to significantly improve the results of other methods reported in the state-of-the-art. The specific features and the feature reduction method are described below.

3.1. Channel-based features

To gather representative aspects of the EEG recordings from two different perspectives, we have used a combination of time-domain and frequency-domain channel-based features. In the time-domain, we have extracted the Relative Energy (RE) and the Logarithmic Relative Energy (LRE). Despite that these are dependent variables, their different relative scaling allows for gathering different aspects of the signal. In the frequency domain, the Spectral Power (SP) has also been computed. In all cases, features have been calculated at each EEG channel, and for each relevant frequency band, namely α (8-13 Hz), β (14-30 Hz) γ (30-47 Hz) and θ (4-7 Hz).

Assuming that C channels (electrodes) are used, this yields a data set with a total of $12 \cdot C$ features. Out of these, $8 \cdot C$ are energy-based (4 frequency bands $\times$ 2 features per band $\times$ C electrodes), and $4 \cdot C$ are frequency-based (each possible combination of electrode and frequency band). These are computed as follows.

3.1.1. Computation of energy-related features

Let us denote by $x^{c,f}$ the resulting signal of length l obtained after filtering the raw EEG output obtained at channel c in the frequency band f (with $c = 1 \dots C$ and $f \in \{\alpha, \beta, \gamma, \theta\}$). To calculate the RE and the LRE, the energies at each frequency band are first computed for each channel as:

$$E(x^{c,f}) = \sum_{i=1}^{l} (x_i^{c,f})^2 \tag{1}$$

with $x_i^{c,f}$ the i-th element of signal $x^{c,f}$.

The Relative Energy (RE) of the signal is then measured with respect to the power of the rest of the frequency bands as:

$$RE(x^{c,f}) = \frac{E(x^{c,f})}{E(x^{c,\alpha}) + E(x^{c,\beta}) + E(x^{c,\gamma}) + E(x^{c,\theta})} \tag{2}$$

Finally, the LRE is then calculated as

$$LRE(x^{c,f}) = log(RE(x^{c,f})) \tag{3}$$

3.1.2. Computation of SP

The Spectral Power (SP) of each signal $x^{c,f}$ has been computed by using a Hamming window with a size of 128 samples, according to the expression:

$$SP(x^{c,f}) = \log\left(\sum_{i=a}^{b} (X_i^{c,f})^2\right) \tag{4}$$

where $X_i^{c,f} = \mathfrak{F}(x_i^{c,f})$, $\mathfrak{F}$ denotes the DFT operator and a y b correspond to the bins in the DFT spectrum that delimit each of the four frequency bands considered.

3.2. Connectivity Features

Connectivity features can also be used to describe brain activities, and reflect the interaction between two cortical areas during an experiment. A connectivity magnitude can be considered between any two channels, leading to $C \cdot (C-1)/2$ features, which correspond to the upper triangle of the $C \times C$ matrix that contains the connectivity values for each channel pair.

A typical connectivity feature is the the mutual information, that we have used in this work and determines how informative a random variable is with respect to another. This has also been successfully used in other previous works [11] to evaluate the relation between the brain activity measured at any two EEG channels, and also identified as the most informative connectivity measure.

The mutual information between two signals x and y is defined in terms of entropy as:

$$I(x; y) = H(x) - H(x|y) \tag{5}$$

where H stands for the entropy of the signal, which can be computed as:

$$H(x) = -\sum p_i * \log p_i \tag{6}$$

with p_i the probability of the i-th element of the time-series x. This yields the expression

$$I(x,y) = -\sum p_{ij}^{XY} \cdot \log\left(\frac{p_{ij}^{XY}}{p_i^X p_j^Y}\right) \quad (7)$$

where p_{ij}^{XY} represents the joint probability of the i-th element of the time-series x and the j-th element of time-series y.

As argued in [32], the computation of entropies for continuous or ordinal data is highly non-trivial, and requires an assumed model of the underlying distributions. To simplify computation and keep consistency with the method used in [11], data has been discretized by using the $\lfloor\cdot\rfloor$ function, that converts each data sample to the highest integer lower than the number. This is equivalent to using an estimator based on a histogram of a fix width of 1.

3.3. Dimensionality reduction scheme

The feature computation presented above yields a total of $12 \cdot C + C \cdot (C-1)/2$ features, an expression that shows a quadratic dependence of the number of channels C. For a typical 32 channel setting, this implies 880 features. To make the problem tractable from a classification perspective, we reduce the number of dimensions by using a novel feature reduction method that is based on the combination of a Welch's t-test with a Principal Component Analysis (PCA).

The Welch's t-test is able to determine if two sets of data are significantly different from each other, by providing an indication of how much separation is present between two groups of data (classes). It can hence be used as a supervised method for feature reduction, by computing its value for all features and discarding the ones that are less useful for classification. To this end, all features where the t-test yields a p-value above an adaptive threshold are eliminated. The remaining ones are retained for further analysis. This first stage of the feature reduction has been inspired by the ANOVA-based method used in other multiple-class problems e.g. [33, 34].

This supervised step is followed by a second unsupervised stage, in which a PCA is run to convert the resulting set of features into a new set in which they are linearly uncorrelated. The number of components retained are dynamically chosen, based on a grid search that is part of the classification process. This steps further reduces the dimensionality of the data, whilst retaining most of its variance.

4. Experimental Results

In order to exhaustively validate the performance and assess the relative merits of the proposal, we have run a large battery of experiments and compared the results to the ones obtained by using other existing and well-accepted methods in the literature.

4.1. Database

One major difficulty associated with the design of classification approaches to process EEG data is the need for a sufficiently large dataset that allows for an appropriate training of the models, and also for a fair comparison to other existing approaches reported in the literature. This has led many authors to create their own dataset, that in many cases was exclusively used for that research and not made publicly available.

A major contribution in this direction is DEAP [9], an open access physiological recordings database specifically created for the analysis of human affective states. The DEAP dataset contains physiological recordings from 32 healthy subjects (50% male, 50% female) aged between 19 and 37 (mean 26.9 years), while watching 40 music videos of 63 seconds each. These videos were selected in order to elicit emotions in each of the 4 quadrants of Russell's Circumplex Model [4], and the emotion was also validated with an online survey. After the video, the subject was asked to report the emotion using Self Assessment Manikin [35] in a range from 1 to 9. EEG signals were recorded at a sampling rate of 512 Hz, using 32 active *AgCl* electrodes placed according to the international 10-20 system. After artifact removal, all signals were down-sampled to 128 Hz.

4.2. Competing methods

As part of the evaluation, we compare the classification performance of the proposal to two other state-of-the-art methods recently proposed in the literature:

- As a first technique, we consider the one that is provided along with the data in DEAP [9]. This method uses the spectral power for five different bands and the spectral power asymmetry between 12 pairs of electrodes. Hence, we will refer to it as **Spectral Power** in this experimental section.

- As a second method, we consider the method presented in [11], and denote it as **Mutual Information**. In this work, authors test the performance of a number of connectivity features, also using the DEAP database. In particular, the performance of Pearson correlation, phase coherence and mutual information were investigated. Performance is reported for the filtered preprocessed signals in each of the four different bands considered in the proposed method ($\alpha, \beta, \gamma, \theta$), and also for the unfiltered preprocessed EEG signals. We have used the combination that leads to the best results, namely the mutual information on all bands.

The two methods use different features but share the same feature reduction method, that differs from the proposal. They use the Fisher linear discriminant [36], which for a given feature f is defined as:

$$J(f) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2} \quad (8)$$

where μ_i and σ_i denote the mean and standard deviation in the two different classes, respectively. The value of the discriminant is calculated for each feature, and an empirically determined threshold (0.3) is applied to discard the less discriminative ones.

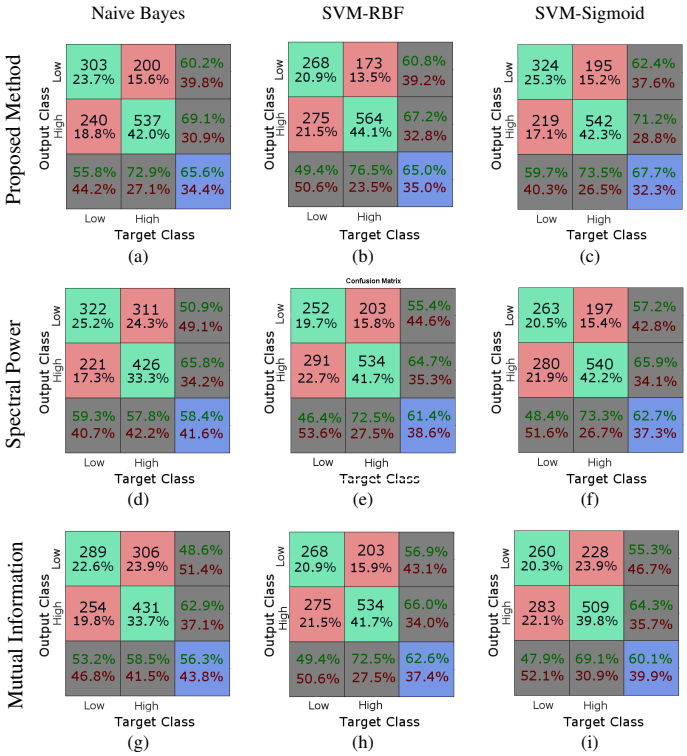

Figure 1: Confusion Matrices for arousal classification performance.

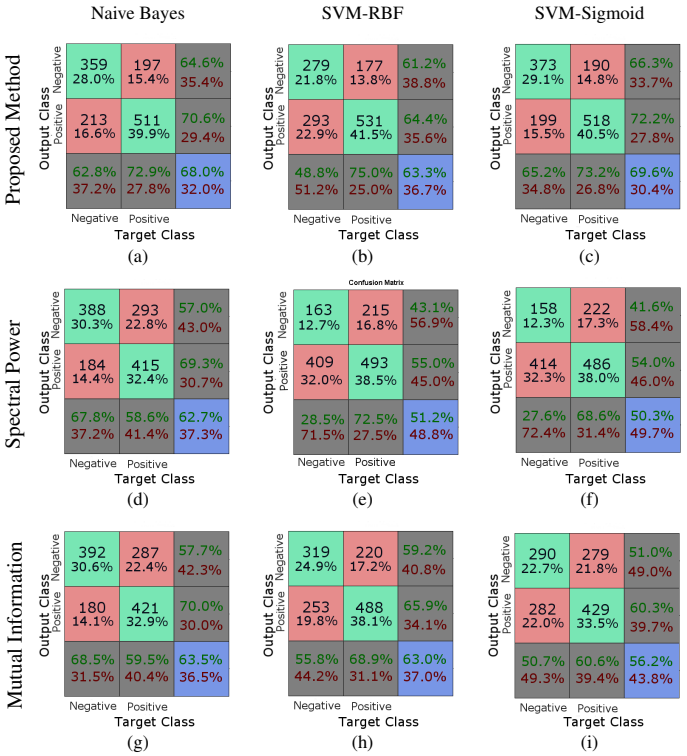

Figure 2: Confusion Matrices for valence classification performance.

The *Spectral Power* and *Mutual Information* methods also use a different classification algorithm. *Spectral Power* uses Naive Bayes, and *Mutual Information* uses a non-linear Support Vector Machine (SVM) with an Radial Basis Function (RBF) kernel. To focus on the effect of the feature selection and reduction mechanism and isolate it from the effect of the classifier, we have conveniently created three variations of each algorithm being compared, each using a different classification technique (Naive Bayes, and a SVM both using a RBF and a sigmoid kernel). In addition, and to ensure that the implementations correspond to the ones reported by the authors in their respective publications, we have requested information to the authors and used the same algorithms. For example, the mutual information in [11] has been computed by using the toolbook described in [37].

4.3. Experimental Setting

In order to compare all algorithms under the same setting, we have adopted a similar experimental setting as in [9], posing two binary classification problems, one for arousal and the other one for valence. To this end, the ratings reported by the user have been used as the ground truth and converted into categorical variables (classes) with two possible values, namely low/high in arousal and positive/negative in valence. On the 9-point rating scales, the threshold was simply placed in the middle. Such a conversion leads to two unbalanced classification problems. 59% of the videos have a high arousal and 57% were rated with a positive valence.

As in [9, 11], we have used a leave-one-out cross validation scheme to evaluate the performance of each competing method. At each step of the cross validation, a single video was used as the test and the rest of the videos for the same subject were used for training. This yields a total of 1 280 steps, each with a training size of 39 and a test size of 1. When required, parameter tuning was done separately at each step by running a grid search in a 5-fold cross validation setting. The combination that led to the best accuracy on average was chosen. For the proposal, the number of dimensions retained by the PCA was built into the grid search as one extra parameter when using a SVM; and determined by cross validation when using Naive Bayes. The threshold used by the Welch's t-test feature selection was increased in steps of 0.01 until the number of dimensions that remained was strictly greater.

Two different measures are provided for comparison. On the one hand, the confusion matrices facilitate interpretation and allow for a visual inspection. On the other, the weighted F1-score summarizes the result of each algorithm under a single number that takes the class balance into consideration. In addition, a statistical analysis has been carried out to test the significance of the results, and box plots have been used as a means to provide a visual cue of the improvement achieved.

All experiments have been run in a Matlab R2014a environment, using Matlab's own implementation of Naive Bayes and the libSVM interface[38] for the SVM implementations.

4.4. Results

As a first result, the classification performance of all algorithms are presented in Figures 1 (arousal) and 2 (valence), by means of the corresponding confusion matrices. In these figures, each row presents the results for each of the three meth-

	Naive Bayes			SVM-RBF			SVM-Sigmoid		
	Mean	Std	F1-Score	Mean	Std	F1-Score	Mean	Std	F1-Score
Proposed Method	**0.656**	(0.116)	**0.644**	**0.650**	(0.132)	**0.630**	**0.677**	**(0.113)**	**0.667**
Spectral Power	0.584	**(0.108)**	0.582	0.614	(0.161)	0.594	0.627	(0.151)	0.609
Mutual Information	0.563	(0.134)	0.557	0.626	**(0.126)**	0.610	0.601	(0.128)	0.586

Table 1: Average accuracy and F1-score for Arousal

	Naive Bayes			SVM-RBF			SVM-Sigmoid		
	Mean	Std	F1-Score	Mean	Std	F1-Score	Mean	Std	F1-Score
Proposed Method	**0.680**	**(0.083)**	**0.675**	**0.633**	(**0.138**)	0.618	**0.696**	**(0.093)**	**0.692**
Spectral Power	0.627	(0.118)	0.624	0.512	(0.235)	0.478	0.503	(0.228)	0.468
Mutual Information	0.635	(0.110)	0.635	0.630	(0.140)	**0.624**	0.562	(0.158)	0.557

Table 2: Average accuracy and F1-score for Valence

ods being compared. Subfigures (a)-(c) correspond to the proposal; (d)-(f) to *Spectral Power*; and (f)-(i) to *Mutual Information*. Each column refers to a different classification method. The first column refers to a Naive Bayes classifier; the second to a SVM using a RBF kernel; and the third to a SVM with a sigmoid kernel. In each subfigure, the first two diagonal (green) cells show the number and percentage of correct classifications for each class. Wrong classifications are shown in the red cells. The number and percentage of correct predictions are shown at the top right corner for the low (negative) class; and just below for the high (positive) class. The number and percentage of instances that are correctly predicted are provided at the bottom left corner for the low (negative) class; and for the high (positive) class at its right side. The overall classification performance (accuracy) is shown at the bottom-right corner. The average accuracy, its standard deviation when results are grouped by users, and the weighted F1-score computed on the entire confusion matrix are reported in Tables 1 and 2 for arousal and valence, respectively. To ease interpretation of the results, highest values for the mean and the F1-score and lowest value for the standard deviation have been printed in bold.

As a first observation, gains in accuracy and F1-score are consistently obtained, with independence from the classification method. This supports the idea that the improvements are a consequence of the combination of different types of features and the hybrid reduction method proposed. The proposal works best when it is integrated with a SVM with a sigmoid kernel, despite that the Naive Bayes also yields a very close performance.

With regard to arousal, our proposal with the worst classifier outperforms the best combination in the other two competing methods. The differences are specially relevant when using Naive Bayes or a SVM with a sigmoid kernel as the classification method. In the first case, the accuracy of the proposal (65.6%) is significantly higher than the one by the other two competing methods (58.4% and 56.3%). When using a SVM with a sigmoid kernel, the accuracy of the proposal reaches 67.7%, in contrast to the 62.7% offered by the best of the other two methods. Best results for *Mutual Information* are obtained when using a SVM classifier with a RBF kernel, that is the classifier selected in the original publication [11]. Results in terms of the F1-score also confirm the superior results obtained with our method. The three best F1-scores correspond to the proposal, and again benefits are specially noticeable when using a SVM with a sigmoid kernel. Another relevant aspect in favor of the proposal relates to the value of the standard deviation. Lower values indicate a higher robustness, with less variation in performance across different users. When combining the methods with a SVM using a sigmoid kernel our proposal not only scores best, but it also yields the lowest standard deviation.

In valence, the proposal presents an even higher accuracy, reaching a 69.6% when using a SVM with a sigmoid kernel and a 68% when using Naive Bayes. The next best performance is for *Mutual Information* using Naive Bayes, which is far from the one obtained with the proposal. The *Spectral Power* method only leads to competitive result when using a Naive Bayes classifier, which is the one used in the original publication [9]. When using a SVM, the predictions of the low class present a performance very close 50%, indicating that the method fails is nearly half the predictions. The best accuracy of *Spectral Power* is close to that of *Mutual Information*, but far from the one achieved with the proposal. Looking at the F1-score values presented in Table 2, we can withdraw similar conclusions. Our proposal outperforms all other combinations when using either Naive Bayes or a SVM with a sigmoid kernel. The F1-scores obtained in these cases are 0.675 and 0.692, respectively. These are significantly higher than the F1-score obtained for the combination of *Mutual Information* with a Naive Bayes classifier (0.635), which is the next best. Only in the case of using a SVM with an RBF kernel, our proposal performs performs slightly worse than *Mutual Information* in terms of the F1-score. As in arousal, we can also observe the generally lower values of the standard deviation. In this occasion, the proposal has the lowest value in all cases, suggesting a higher robustness of the approach.

These results are accompanied by a significance test, that compares the mean values of the accuracy to ascertain that differences in favor of the method proposed (using a SVM with a sigmoid kernel) are significant from a statistical point of view. The results for a paired t-test have been complemented with a Wilcoxon signed-rank test (single-tailed). This is because of the presence of outliers, and also because the assumption that the differences between pairs are normally distributed in the t-test cannot be guaranteed. To perform this test, we have grouped the classification results per user, so that each value represents

		Spectral Power			Mutual Information		
		Naive Bayes	SVM-RBF	SVM-Sigmoid	Naive Bayes	SVM-RBF	SVM-Sigmoid
Arousal	t-test	0.0003	0.0198	0.0486	$< 10^{-4}$	0.0185	0.0013
	Wilcoxon's	0.0004	0.0178	0.0487	$< 10^{-4}$	0.0194	0.0022
Valence	t-test	0.0028	0.0004	0.0002	0.0118	0.0001	$< 10^{-4}$
	Wilcoxon's	0.0029	$< 10^{-4}$	$< 10^{-4}$	0.0134	0.0003	$< 10^{-4}$

Table 3: Significance paired tests between our proposal using a SVM with a sigmoid kernel and the rest of the methods.

the accuracy obtained with a method on the 40 videos for a particular subject. This yields a set of 32 measurement for each method, and allows taking our method as a reference and compare it against each of the competing alternatives, both in arousal and valence terms. Significance results are presented in Table 3, using the p-value. The p-values provided by both statistical tests are very close. The improvements obtained with the proposal are statistically significant in all cases.

Figures 3 and 4 also help observe the relatively higher performance in accuracy offered by the proposal in comparison to the rest of the approaches, when using the classification mechanism specified in their respective publications. In this figure, the results obtained for each user have been grouped and depicted graphically by using notched box plots. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25^{th} and 75^{th} percentiles, respectively. The whiskers extend to an interval whose extremes are $m - 1.57(q_3 - q_1)/\sqrt{(n)}$ and $m + 1.57(q_3 - q_1)/\sqrt{(n)}$, where m is the median, and q_1 and q_3 are the 25th and 75th percentiles. Samples outside the whiskers are considered outliers and plotted individually using red crosses.

With regard to arousal, the box plot in Figure 3 reveals two outliers in the case of the proposal. This indicates that there are two users for whom the approach did not work well. The relatively low values of these outliers negatively affect the average accuracy, that is still higher than for the other two methods, as it was shown in Table 1. This supports the better performance of the proposal in the general case. The shorter box for the proposal also indicates a higher robustness, with more stable results. In addition, the notch that corresponds to the proposal does not overlap with the median of any other, a fact that confirms our findings in the significance analysis. The box plot in Figure 4 also reveals large differences in the median values of the valence. Despite the outliers that are in the proposal and in *Spectral Power*, results for *Mutual Information* present a higher variability and hence a longer box.

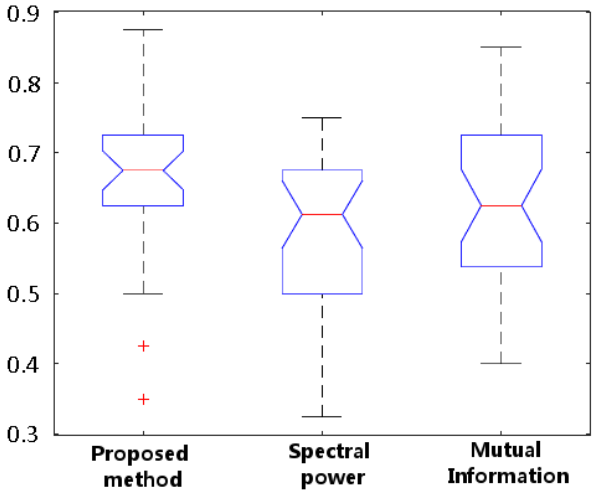

Figure 3: Result comparison for Arousal.

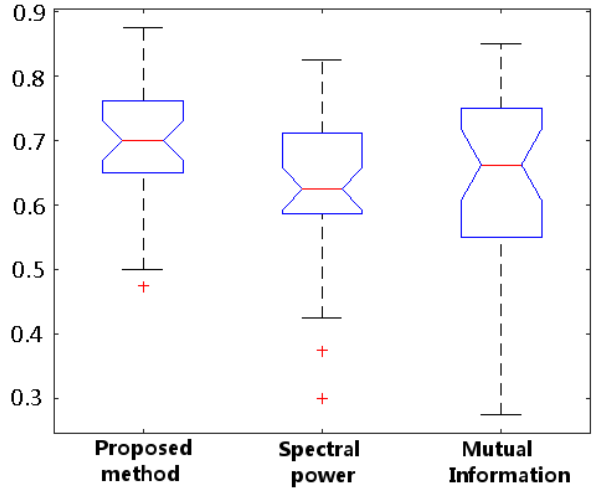

Figure 4: Result comparison for Valence.

5. Conclusions

The use of EEG signals for emotion recognition is a relatively recent research field. Both channel-based and connectivity-based feature sets have been commonly used in the literature, and several works have already addressed their suitability to emotion recognition, by analyzing their performance when they are individually used [6, 13, 11]. However, the combination of features has been less investigated in this context.

An inherent problem to using a diversity of features is related to the dimensionality of the resulting representation. A channel-based measure produces one feature per channel (electrode), while a connectivity-based one yields a number of features that grows quadratically with the number of channels. When combination approaches are used, this may easily lead to a set composed of several hundreds of features. While these contain a large amount of information that can be exploited in a machine learning context, the low number of examples generally available makes them poorly suited for classification algorithms without a feature reduction scheme. While combining different types of features increases the amount of information, they also increase estimation errors. Hence, the feature increase does not necessarily have a positive influence in the classification results. Finding the most adequate balance between the amount of information and the data dimensionality is a major challenge in this type of problem. In the proposal presented in this paper, we have fused a large number of features (880) of different nature. To compensate for the dimensionality increase, we have designed an adaptive feature reduction methods that combines supervised and unsupervised techniques to yield a drastic reduction. The approach has proven successful as compared to other state-of-the-art methods recently published, independent of the classification method used. Best resutls have been observed when combined with a SVM using a sigmoid kernel.

Despite the relatively high accuracy obtained in the arousal and valence dimensions, the binary decisions on these variables are only able to locate the emotion in one of the four quadrants of existing 2-D models e.g. Russell's Circumplex Model [4]. When arousal and valence are used to predict the subject's affective state in the form of a non-binary output e.g. a concrete basic emotion, the problem turns into a multi-class classification one. In this case, the use of 2-D emotional models to combine the arousal and valence levels into a single output generally lacks of the precision required, because errors in the two variables accumulate. Therefore, adapting the proposal to a multi-class setting is a natural extension of the work presented. So it is the use of regression methods that are able to quantify the arousal and valence levels to locate the emotion more accurately in a 2-D space. These extensions would contribute to facilitating a seamless integration of the emotion recognition system within existing applications, including recommender systems and Intelligent Tutors e.g. [39].

The positive results obtained at the problem at hand reinforce the potential benefits of gathering diverse types of features in the representation of EEG signals. They also outline the importance of combining this type of strategies with adequate feature reduction methods to deal with the high dimensionality. In this paper, we have focused on different types of connectivity and channel-based features, extracted in both the temporal and frequency domain. However, other current trend not explored in this paper is the use of unsupervised learning to compute new features from existing ones. Some examples of this are the approach proposed by [40] for intrusion detection or the more general method presented in [41], that expand the original feature vectors by computing new features using distances from each data sample to a number of centroids found by a clustering algorithm. In this same line, the use of distance-based features to a set of reference patterns, or the related concept of pairwise dissimilarities [42], could effectively be used to enrich or replace the information provided in the original feature vectors.

Acknowledgments

This work has been partly supported by UWS Vice Principal Fund and Spanish Ministry of Economy and Competitiveness through project TIN2014-59641-C2-1-P.

References

[1] R. Picard, Affective Computing, MIT Press, 2000.
[2] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (1992) 169–200.
[3] R. Plutchick, The nature of emotions, American Scientist 89 (2001) 344–350.
[4] J. A. Russell, Affective space is bipolar, Journal of Personality and Social Psychology (1979).
[5] A. Mehrabian, Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament, Current Psychology 14 (1996) 261–292.
[6] M.-K. Kim, M. Kim, E. Oh, S.-P. Kim, A review on the computational methods for emotional state estimation from the human EEG, Computational and mathematical methods in medicine (2013).
[7] Y. Liu, O. Sourina, EEG-based subject-dependent emotion recognition algorithm using fractal dimension, in: IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 3166–3171.
[8] Y. Liu, O. Sourina, M. K. Nguyen, Real-time EEG-based human emotion recognition and visualization, in: International Conference on Cyberworlds (CW), pp. 262–269.
[9] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A database for emotion analysis using physiological signals, IEEE Transactions on Affective Computing 3 (2012) 18–31.
[10] P. Petrantonakis, L. Hadjileontiadis, Adaptive emotional information retrieval from EEG signals in the time-frequency domain, IEEE Transactions on Signal Processing 60 (2012) 2604–2616.
[11] M. Chen, J. Han, L. Guo, J. Wang, I. Patras, Identifying valence and arousal levels via connectivity between EEG channels, in: IEEE International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 63–69.
[12] P. Arnau-González, N. Ramzan, M. Arevalillo-Herrráez, A method to identify affect levels from EEG signals using two dimensional emotional models, in: 30th European Simulation and Modelling Conference (ESM'2016), pp. 1–5.
[13] R. Jenke, A. Peer, M. Buss, Feature extraction and selection for emotion recognition from EEG, IEEE Transactions on Affective Computing 5 (2014) 327–339.
[14] S. Hadjidimitriou, V. Charisis, L. Hadjileontiadis, Towards a practical subject-independent affective state recognition based on time-domain EEG feature extraction, International Journal of Heritage in the Digital Era 4 (2015) 165–178.
[15] R. Gupta, T. H. Falk, et al., Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization, Neurocomputing 174 (2016) 875–884.
[16] V. Bono, D. Biswas, S. Das, K. Maharatna, Classifying human emotional states using wireless EEG based ERP and functional connectivity measures, in: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 200–203.
[17] Y.-Y. Lee, H. S, Classifying different emotional states by means of EEG-based functional connectivity patterns, PLoS ONE 9 (2014).
[18] P. Petrantonakis, L. Hadjileontiadis, Emotion recognition from EEG using higher order crossings, IEEE Transactions on Information Technology in Biomedicine 14 (2010) 186–197.
[19] B. Kedem, S. Yakowitz, Time series analysis by higher order crossings, IEEE press New York, 1994.

[20] Q. Wang, O. Sourina, Real-time mental arithmetic task recognition from EEG signals, IEEE Transactions on Neural Systems and Rehabilitation Engineering 21 (2013) 225–232.
[21] L. Bozhkov, P. Koprinkova-Hristova, P. Georgieva, Reservoir computing for emotion valence discrimination from {EEG} signals, Neurocomputing (2016) –.
[22] K. Pearson, Note on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London 58 (1895) 240–242.
[23] A. Pikovsky, M. Rosenblum, J. Kurths, Synchronization: a universal concept in nonlinear sciences, volume 12, Cambridge university press, 2003.
[24] M. Hassan, O. Dufor, I. Merlet, C. Berrou, F. Wendling, EEG source connectivity analysis: from dense array recordings to brain networks, PloS one 9 (2014) e105041.
[25] G. Lee, M. Kwon, S. K. Sri, M. Lee, Emotion recognition based on 3d fuzzy visual and eeg features in movie clips, Neurocomputing 144 (2014) 560–568.
[26] Q. Zhang, S. Jeong, M. Lee, Autonomous emotion development using incremental modified adaptive neuro-fuzzy inference system, Neurocomputing 86 (2012) 33 – 44.
[27] M. Soleymani, S. Asghari-Esfeden, M. Pantic, Y. Fu, Continuous emotion detection using EEG signals and facial expressions, in: IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6.
[28] T. Baltrusaitis, N. Banda, P. Robinson, Dimensional affect recognition using continuous conditional random fields, in: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, pp. 1–8.
[29] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[30] M. Soleymani, S. Asghari-Esfeden, Y. Fu, M. Pantic, Analysis of EEG signals and facial expressions for continuous emotion detection, IEEE Trans. Affective Computing 7 (2016) 17–28.
[31] G. V. Trunk, A problem of dimensionality: A simple example, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (1979) 306–307.
[32] G. Brown, A. Pocock, M.-J. Zhao, M. Luján, Conditional likelihood maximisation: A unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.
[33] K. J. Johnson, R. E. Synovec, Pattern recognition of jet fuels: comprehensive gc× gc with anova-based feature selection and principal component analysis, Chemometrics and Intelligent Laboratory Systems 60 (2002) 225–237.
[34] K. M. Pierce, J. L. Hope, K. J. Johnson, B. W. Wright, R. E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, Journal of Chromatography A 1096 (2005) 101–110.
[35] M. M. Bradley, P. J. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, Journal of behavior therapy and experimental psychiatry 25 (1994) 49–59.
[36] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188.
[37] R. Moddemeijer, On estimation of entropy and mutual information of continuous distributions, Signal Processing 16 (1989) 233–246.
[38] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27.
[39] D. Arnau, M. Arevalillo-Herráez, J. A. González-Calero, Emulating human supervision in an intelligent tutoring system for arithmetical problem solving, IEEE Transactions in Learning Technologies 7 (2014) 155–164.
[40] C. Guo, Y. Zhou, Y. Ping, Z. Zhang, G. Liu, Y. Yang, A distance sum-based hybrid method for intrusion detection., Applied Intelligence 40 (2014) 178–188.
[41] C.-F. Tsai, W.-Y. Lin, Z.-F. Hong, C.-Y. Hsieh, Distance-based features in pattern classification., EURASIP J. Adv. Sig. Proc. 2011 (2011) 62.
[42] R. P. W. Duin, M. Loog, E. Pkalska, D. M. J. Tax, Feature-Based Dissimilarity Space Classification, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 46–55.